Análise de Dados

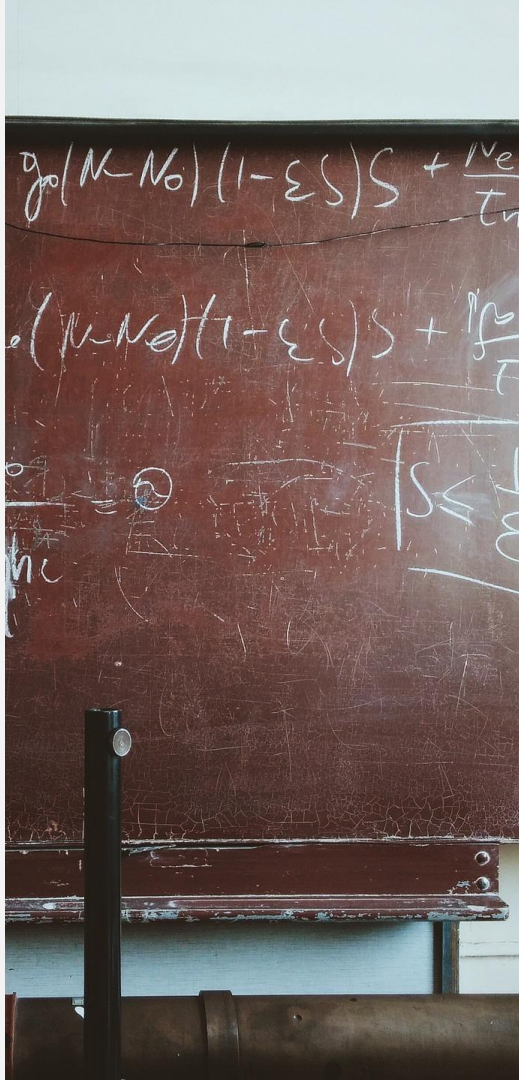
Estatística Inferencial

CONCEITOS ABORDADOS

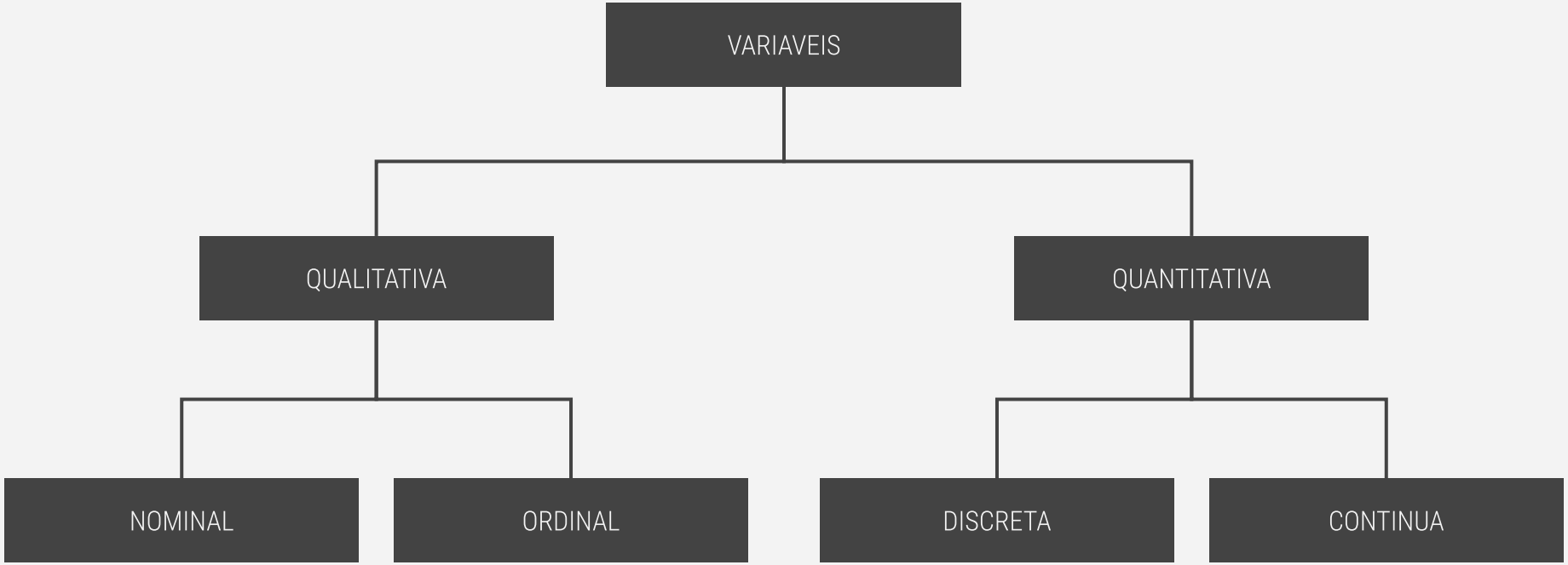
- Variáveis Aleatórias
- População e Amostra
- Modelos Probabilisticos
  - Discretos
  - Continuous
- Teorema do Limite Central

01

ANÁLISE BIVARIADA



TIPOS DE VARIAVEIS



ANÁLISE BIVARIADA

Sempre que vamos realizar uma análise bivariada temos 3 situações:

As duas variáveis são qualitativas	As duas variáveis são quantitativas	Uma variável qualitativa e outra quantitativa

Qualitativas

QUALITATIVAS

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas, isto é, queremos conhecer o grau de dependência entre elas, de modo que possamos prever melhor o resultado de uma delas quando conhecermos a realização da outra

GENERO

CURSO

QUALITATIVAS

TABELA DE CONTINGÊNCIA

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	40 (40%)	200 (67%)	240 (60%)
Mulheres	60 (60%)	100 (33%)	160 (40%)
total	100 (100%)	300 (100%)	400 (100%)

TABELA DE CONTINGÊNCIA

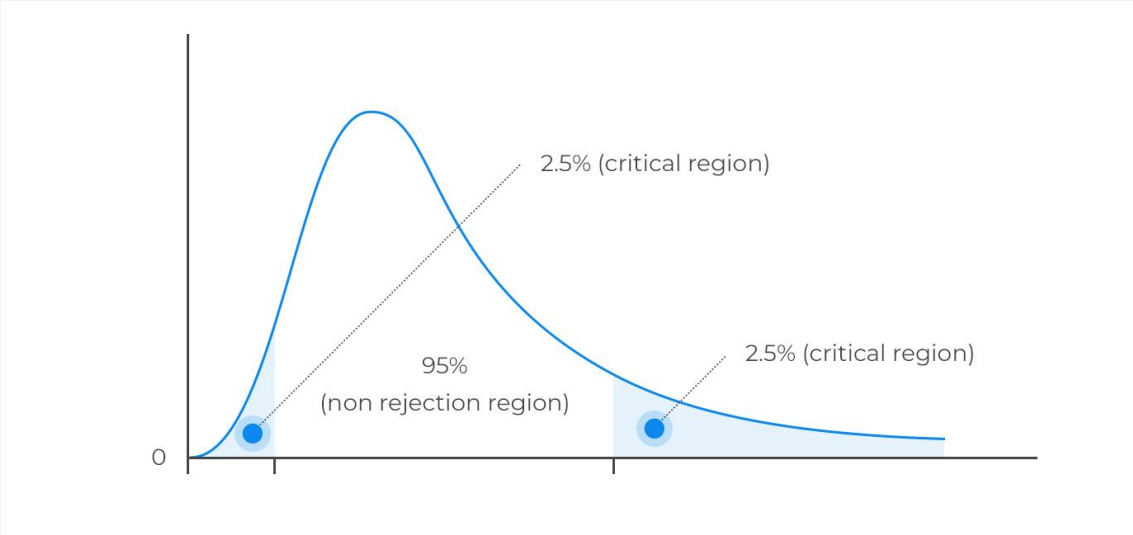
Será que o sexo tem dependência com o tipo do curso?

Como podemos saber?

sexo	Curso 1 Estatística	Curso 2 Engenharia	total
Homens	40 (40%)	200 (67%)	240 (60%)
Mulheres	60 (60%)	100 (33%)	160 (40%)
total	100 (100%)	300 (100%)	400 (100%)

TESTE CHI-QUADRADO

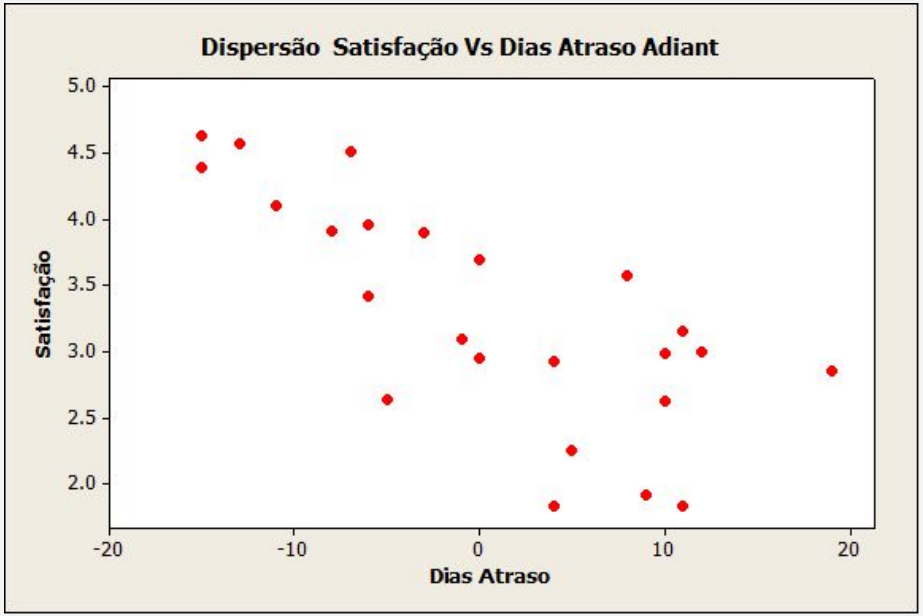
De modo geral, a quantificação do grau de associação entre duas variáveis pode ser feita através do teste de chi-quadrado, onde passamos o p-value e o graus de liberdade da nossa análise.



Quantitativa

QUALITATIVAS

Podemos verificar a dependência de duas variáveis quantitativas, apenas visualizando um gráfico de dispersão.



QUALITATIVAS

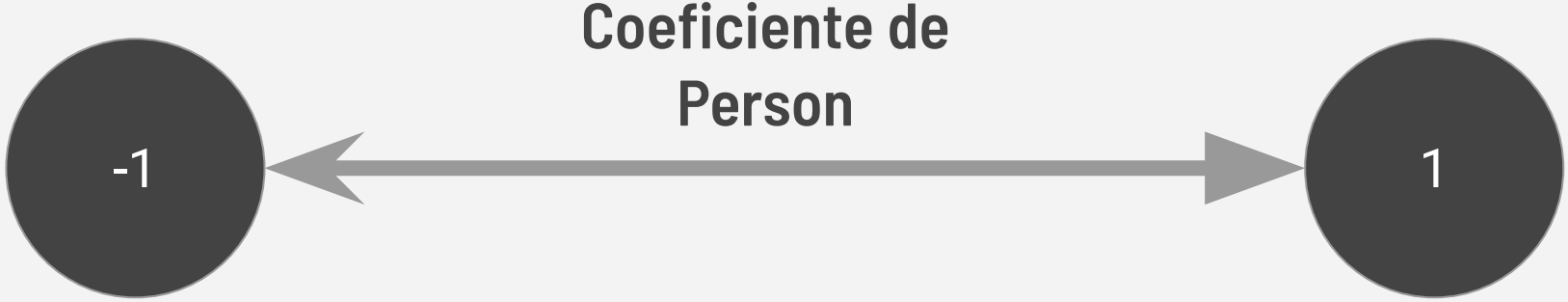
Porém como podemos quantificar essa dependência?

QUALITATIVAS

Podemos utilizar um coeficiente de correlação entre duas variáveis é uma medida do grau de associação entre elas e também da proximidade dos dados a uma reta

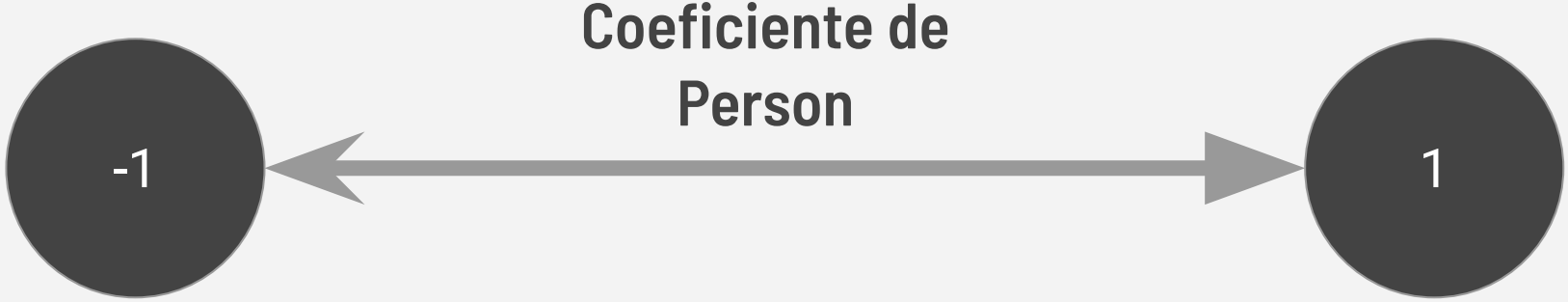
QUALITATIVAS

Podemos utilizar um coeficiente de correlação entre duas variáveis é uma medida do grau de associação entre elas e também da proximidade dos dados a uma reta

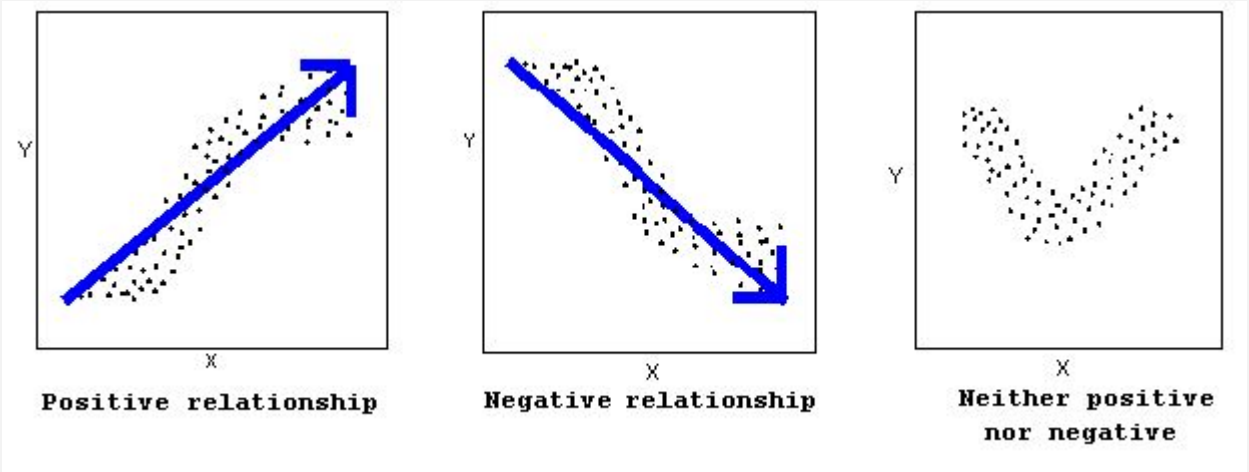


QUALITATIVAS

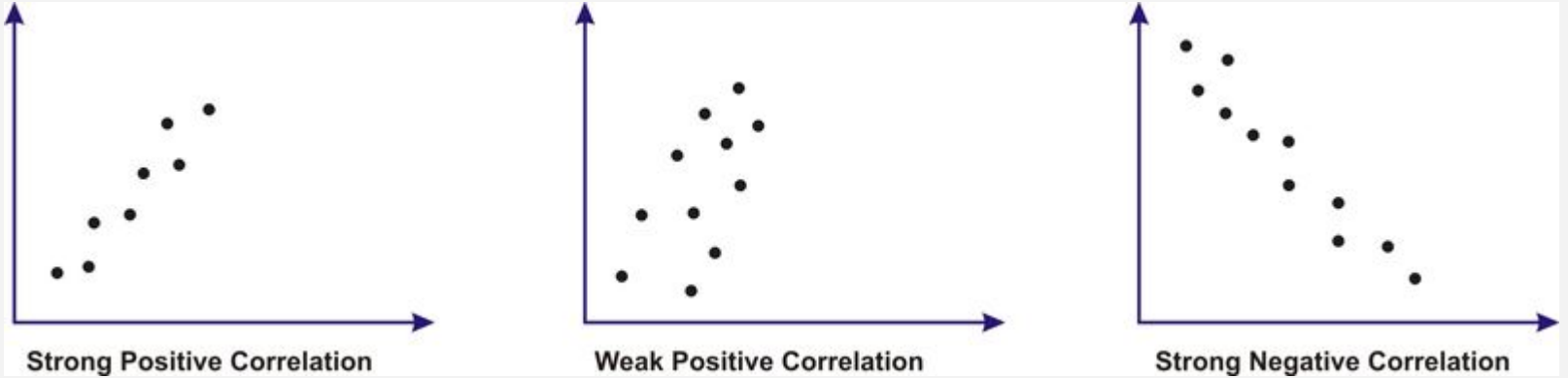
Podemos utilizar um coeficiente de correlação entre duas variáveis é uma medida do grau de associação entre elas e também da proximidade dos dados a uma reta



DIREÇÃO


Y
X
Positive relationship
Y
X
Negative relationship
Y
X
Neither positive
nor negative

FORÇA


Strong Positive Correlation
Weak Positive Correlation
Strong Negative Correlation

Transformações

TRANSFORMAÇÃO

A transformação dos dados é uma prática para evitar que nossas análises seja impactadas pelas as diferenças de escala dos dados.

NORMALIZAÇÃO

A normalização é uma técnica de dimensionamento na qual os valores são deslocados e redimensionados para que eles acabem variando entre 0 e 1. Também é conhecido como escala Min-Max.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Pontos Importantes

- Quando o valor de X é o valor mínimo na coluna, o numerador será 0, e portanto X' é 0
- Por outro lado, quando o valor de X é o valor máximo na coluna, o numerador é igual ao denominador e, portanto, o valor de X' é 1
- Se o valor de X estiver entre o valor mínimo e o máximo, então o valor de X' está entre 0 e 1

PADRONIZAÇÃO

A padronização é outra técnica de dimensionamento onde os valores estão centrados em torno da média com um desvio padrão unitário. Isso significa que a média do atributo torna-se zero e a distribuição resultante tem um desvio padrão unitário.

$$X' = \frac{X - \mu}{\sigma}$$

Pontos Importantes

- Quando o valor de X é o valor mínimo na coluna, o numerador será 0, e portanto X' é 0
- Por outro lado, quando o valor de X é o valor máximo na coluna, o numerador é igual ao denominador e, portanto, o valor de X' é 1
- Se o valor de X estiver entre o valor mínimo e o máximo, então o valor de X' está entre 0 e 1

Qual utilizar?

- A normalização é boa de usar quando você sabe que a distribuição de seus dados não segue uma distribuição gaussiana. Isso pode ser útil em algoritmos que não assumem qualquer distribuição dos dados como K-Nearest Neighbors e Neural Networks.
- A padronização, por outro lado, pode ser útil nos casos em que os dados seguem uma distribuição gaussiana. No entanto, isso não precisa ser necessariamente verdade. Além disso, ao contrário da normalização, a padronização não tem um alcance delimitante. Assim, mesmo que você tenha outliers em seus dados, eles não serão afetados pela padronização.